

SortMeRNA User Manual

Evguenia Kopylova
evguenia.kopylova@lifl.fr

April 2013, version 1.7

Contents

1	Introduction	3
2	Installation	3
2.1	Install from source code	3
2.2	Install from precompiled code	5
2.3	Uninstall	6
3	Databases	6
4	How to run SortMeRNA	7
4.1	Index the rRNA database: command ‘ buildtrie ’	7
4.1.1	Example 1: buildtrie	7
4.2	Filter reads against the indexed rRNA database: command ‘ sortmerna ’	9
4.2.1	Example 2: sortmerna on multiple databases, with output of accepted reads sorted by database	10
4.2.2	Example 3: sortmerna on multiple databases, with output of accepted reads directed to one single file	12
4.2.3	Example 4: sortmerna on paired-end reads (1 input file - paired-end reads are interleaved)	13
4.2.4	Example 5: sortmerna on forward-reverse paired-end reads (2 input files) . .	15
5	SortMeRNA parameters	16

1 Introduction

Copyright (C) 2012-2013 Bonsai Bioinformatics Research Group
(LIFL - Université Lille 1), CNRS UMR 8022, INRIA Nord-Europe

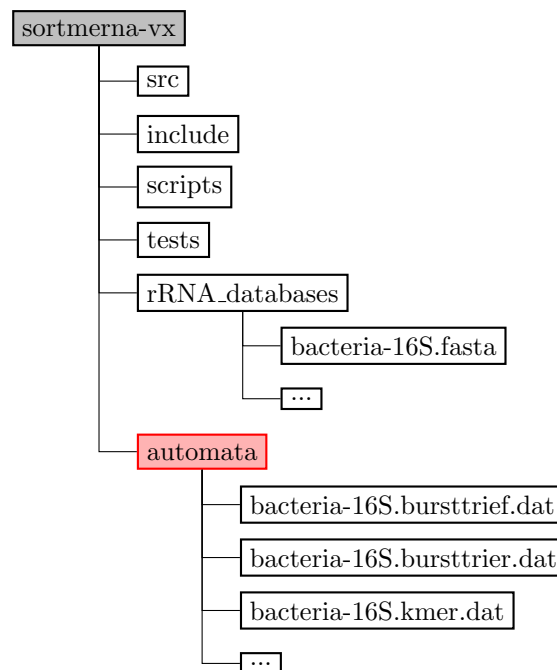
SortMeRNA is a software designed to filter metatranscriptomic reads data. It takes as input a file of reads (fasta or fastq format) and an rRNA database file, and sorts apart the accepted reads and the rejected reads into two files specified by the user.

For questions & help, please contact:

- | | |
|----------------------|---------------------------|
| 1. Evguenia Kopylova | evguenia.kopylova@lifl.fr |
| 2. Laurent Noe | laurent.noe@lifl.fr |
| 3. Helene Touzet | helene.touzet@lifl.fr |

2 Installation

Figure 1: `sortmerna-vx` directory tree



2.1 Install from source code

1. Download `sortmerna-vx.tar.gz` from <http://bioinfo.lifl.fr/RNA/sortmerna>
2. Extract the source code package into a directory of your choice, for quick installation without tests, type:

Quick Installation (no testing)

```
> ./configure
> make
> make install (with root permissions)

Go to Step 5.
```

3. For installation with testing, proceed by typing:

```
> ./configure
> make
> make check
```

4. At this point, two executables `buildtrie` and `sortmerna` will be located in the `sortmerna-vx` directory, as well as the indexed rRNA databases in `sortmerna-vx/automata`. If the user would like to install the executables into `/usr/local/bin` and the `sortmerna-vx` directory into `/usr/local/sortmerna-vx`, then type,

```
> make install (with root permissions)
> make installcheck
```

5. SortMeRNA indexes a database by writing its contents to the `'/automata'` folder (see Figure 1), which is initially located in the directory `'/sortmerna-vx/automata'`. It is essential to set the `$SORTMERNADIR` environmental variable to the path where the `'/automata'` folder is found, so that the program may read and write from it. The user may move the `'/automata'` folder to a workspace with larger memory, separately from the directory `'/sortmerna-vx'`. To find the path of the `'/automata'` folder, go into the directory where it is located (we assume here it is located under the directory `sortmerna-vx`) and type,

```
> pwd
/path/to/sortmerna-vx
```

Open the `~/.bashrc` (or `~/.profile`) file in any editor and add the line,

```
> export SORTMERNADIR="/path/to/sortmerna-vx"
```

If the user installed SortMeRNA in Step (2) or Step (4), go to Step (6). Otherwise, the user must also include the path of the executable files in the `PATH` variable (we assume `buildtrie` and `sortmerna` are located in the `sortmerna-vx` directory),

```
> export PATH="$PATH:/path/to/sortmerna-vx"
```

6. Run the `~/.bashrc` (or `~/.profile`) file to add the variable `$SORTMERNADIR` and update the variable `$PATH` in the list of environment variables,

```
> source ~/.bashrc
or
```

- ```
> source ~/.profile
```
7. Check the path of the executables (if installed manually),
 

```
> which sortmerna buildtrie
/path/to/sortmerna-vx/sortmerna
/path/to/sortmerna-vx/buildtrie
```
  8. Check that \$SORTMERNADIR has been added,
 

```
> echo $SORTMERNADIR
/path/to/sortmerna-vx (if it has not been added,
 this path will be empty)
```
  9. To begin using SortMeRNA, type 'buildtrie -h' or 'sortmerna -h'. If the user installed SortMeRNA in Step 3, then the public rRNA databases distributed with SortMeRNA have been indexed in \$SORTMERNADIR/automata and the user may directly run the command **sortmerna**. Otherwise, the user must firstly index the databases with the command **buildtrie** before they can run the command **sortmerna**.

## 2.2 Install from precompiled code

1. Download the latest binary distribution of SortMeRNA from <http://bioinfo.lifl.fr/RNA/sortmerna>
2. Extract the source code package into a directory of your choice,
 

```
> tar -zxvf sortmerna-vx.tar.gz
> cd sortmerna-vx
```
3. SortMeRNA indexes a database by writing its contents to the '/automata' folder (see Figure 1), which is initially located in the directory '/sortmerna-vx/automata'. It is essential to set the \$SORTMERNADIR environmental variable to the path where the '/automata' folder is found, so that the program may read and write from it. The user may move the '/automata' folder to a workspace with larger memory, separately from the directory '/sortmerna-vx'. To find the path of the '/automata' folder, go into the directory where it is located (we assume here it is located under the directory **sortmerna-vx**) and type,
 

```
> pwd
/path/to/sortmerna-vx
```
4. Open the ~/.bashrc (or ~/.profile) file in any editor and add the line,
 

```
> export SORTMERNADIR="/path/to/sortmerna-vx"
> export PATH="$PATH:/path/to/sortmerna-vx"
```
5. Run the ~/.bashrc (or ~/.profile) file to add the variable \$SORTMERNADIR and update the variable \$PATH in the list of environment variables,

```
> source ~/.bashrc
or
> source ~/.profile
```

6. Check that `$SORTMERNADIR` has been added,

```
> echo $SORTMERNADIR
/path/to/sortmerna-vx (if it has not been added,
 this path will be empty)
```

7. Check that the `$PATH` has been updated with the additional directory search path,

```
> echo $PATH
/usr/local/bin:/usr/bin:...:/path/to/sortmerna-vx
```

8. Check the path of the executables,

```
> which sortmerna buildtrie
/path/to/sortmerna-vx/sortmerna
/path/to/sortmerna-vx/buildtrie
```

9. To begin using SortMeRNA, type ‘`buildtrie -h`’ or ‘`sortmerna -h`’. If the user installed SortMeRNA in Step 3, then the public rRNA databases distributed with SortMeRNA have been indexed in `$SORTMERNADIR/automata` and the user may directly run the command `sortmerna`. Otherwise, the user must firstly index the databases with the command `buildtrie` before they can run the command `sortmerna`.

## 2.3 Uninstall

If the user installed SortMeRNA using the command ‘`make install`’, then they can use the command ‘`make uninstall`’ to uninstall SortMeRNA (with root permissions).

## 3 Databases

SortMeRNA comes prepackaged with 8 databases,

| representative database           | id % | average id % | # seq | origin                 | # seq  | filtered to remove |
|-----------------------------------|------|--------------|-------|------------------------|--------|--------------------|
| silva-bac-16s-database-id85.fasta | 85   | 91.6         | 8174  | SILVA SSU Ref NR v.111 | 244077 | 23s                |
| silva-arc-16s-database-id95.fasta | 95   | 96.7         | 3845  | SILVA SSU Ref NR v.111 | 10919  | 23s                |
| silva-euk-18s-database-id95.fasta | 95   | 96.7         | 4512  | SILVA SSU Ref NR v.111 | 31862  | 26s,28s,23s        |
| silva-bac-23s-database-id95.fasta | 98   | 99.4         | 3055  | SILVA LSU Ref v.111    | 19580  | 16s,26s,28s        |
| silva-arc-23s-database-id95.fasta | 98   | 99.5         | 164   | SILVA LSU Ref v.111    | 405    | 16s,26s,28s        |
| silva-euk-28s-database-id95.fasta | 98   | 99.1         | 4578  | SILVA LSU Ref v.111    | 9321   | 18s                |
| rfam-5s-database-id98.fasta       | 98   | 99.2         | 59513 | RFAM                   | 116760 | –                  |
| rfam-5.8s-database-id98.fasta     | 98   | 98.9         | 13034 | RFAM                   | 225185 | –                  |

The tool UCLUST was used to reduce the size of the original databases.

**id %:** members of the cluster must have identity at least this % id with the representative sequence  
**average id %:** average identity of a cluster member to the representative sequence

**Remark:** The user must first index the fasta database by using the command **buildtrie** and then filter reads against the database using the command **sortmerna**.

## 4 How to run SortMeRNA

### 4.1 Index the rRNA database: command ‘buildtrie’

The executable **buildtrie** indexes an rRNA database.

To see the man page for **buildtrie**,

```
> buildtrie -h
```

This program builds a Burst trie on an input rRNA database file in fasta format and stores the material in binary files under the folder ‘automata’

```
./buildtrie --db [path to rrnas database file name {.fasta}] {OPTIONS}
```

The list of **OPTIONS** can be left blank, the default values will be used:

```
-L length of the sliding window (the seed)
 (default: 18)

-F search only the forward strand

-R search only the reverse-complementary strand
 (default: both strands are searched)

-h help
```

There are eight rRNA representative databases provided in the ‘**sortmerna/rRNA.databases**’ folder. All databases were derived from the SILVA SSU and LSU databases (release 111) and the RFAM databases using the tool UCLUST. Additionally, the user can index their own database.

#### 4.1.1 Example 1: buildtrie

```
> buildtrie --db ~/sortmerna/rRNA_databases/silva-bac-16s-database-id85.fasta
```

```
Burst trie(s) built in: 36.7594s
Writing Burst trie forward to silva-bac-16s-database-id85.bursttrief.dat
Writing Burst trie reverse to silva-bac-16s-database-id85.bursttrier.dat
```

Done.

The indexed databases (ex. `silva-bac-16s-database-id85.bursttrief.dat`) will be stored in the directory `‘/some/path/to/sortmerna/automata’` (the path stored in variable `$SORTMERNADIR`, which was established in Step 1-4 of Subsection 2.2 or Subsection 2.3) and later retrieved by the command `sortmerna`, explained in the following section.

## 4.2 Filter reads against the indexed rRNA database: command ‘sortmerna’

The executable `sortmerna` filters rRNA reads against an indexed rRNA database.

To see the man page for `sortmerna`,

```
> sortmerna -h
```

To run SortMeRNA, type in any order after ‘sortmerna’:

```
--I [illumina reads file name {fasta/fastq}]

--454 [roche 454 reads file name {fasta/fastq}]

-n number of databases to use (must precede --db)

--db [rrnas database name(s)]
 One database,
 ex 1. -n 1 --db /path1/database1.fasta

 Multiple databases,
 ex 2. -n 2 --db /path2/database2.fasta /path3/database3.fasta

{OPTIONS}
```

The list of OPTIONS can be left blank, the default values will be used:

```
--accept [accepted reads file name]
--other [rejected reads file name]
 (default: no output files are created)

--bydbs output the accepted reads by database
 (default: concatenated file of reads)

--log [overall statistics file name]
 (default: no statistics file created)

--paired-in put both paired-end reads into --accept file
--paired-out put both paired-end reads into --other file
 (default: if one read is accepted and the other is not,
 separate the reads into --accept and --other files)

-r ratio of the number of hits on the read / read length
 (default Illumina: 0.25, Roche 454: 0.15)

-F search only the forward strand
-R search only the reverse-complementary strand
 (default: both strands are searched)

-a number of threads to use
```

```

 (default: 1)

-m (m x 4096 bytes) for loading the reads into memory
 ex. '-m 4' means 4*4096 = 16384 bytes will be allocated for the reads
 note: maximum -m is 1020040
 (default: m = 262144 = 1GB)

-v verbose
 (default: deactivated)

-h help

--version version number

```

The command **sortmerna** takes as input a list of rRNA databases (in fasta format) and a set of Illumina or Roche 454 reads (in fasta or fastq format), and filters out the reads matching to at least one of the rRNA databases. The user has an option to output the accepted reads into a single file (default), or into multiple files sorted by the closest matching database (add the flag **--bydbs**). The indexed part of the databases created by **buildtrie** is loaded into **sortmerna** independently.

The user can adjust the amount of memory allocated for loading the reads through the command option **-m**. By default, **-m** is set to be high enough for 1GB. If the reads file is larger than 1GB, then **sortmerna** internally divides the file into partial sections of 1GB and executes one section at a time. Hence, if a user has an input file of 15GB and only 1GB of RAM to store it, the file will be processed in partial sections using **mmap** without having to physically split it prior to execution. Otherwise, the user can set **-m** high enough to store all 15GB in RAM.

#### 4.2.1 Example 2: sortmerna on multiple databases, with output of accepted reads sorted by database

```

> sortmerna -n 3
 --db ~/sortmerna/rRNA_databases/silva-bac-23s-database-id98.fasta
 ~/sortmerna/rRNA_databases/silva-bac-16s-database-id85.fasta
 ~/sortmerna/rRNA_databases/silva-euk-18s-database-id95.fasta
 --accept rrna
 --bydbs
 --454 SRR106861-filtered.fasta
 --log bilan
 -a 3
 -v

```

WARNING: option '--other' has been left blank, no output file for rejected reads ..

```

Welcome to SortMeRNA!
Copyright (C) 2012 Bonsai Bioinformatics Research Group
LIFL, Université Lille 1, CNRS UMR 8022, INRIA 2012

```

The size of the reads file <33862846> bytes  
will be executed in 1 partial section(s) of size  
<33862846> bytes

[Partial section # 1]

-----  
Time to mmap reads and set up pointers: 0.2164s

Begin analysis of: ./rRNA\_databases/silva-bac-23s-database-id98.fasta

Time to load the Burst trie: 1.4488s  
Begin parallel traversal ...  
Time of parallel traversal of automata: 14.2974s

Begin analysis of: ./rRNA\_databases/silva-bac-16s-database-id85.fasta

Time to load the Burst trie: 2.2975s  
Begin parallel traversal ...  
Time of parallel traversal of automata: 17.3430s

Begin analysis of: ./rRNA\_databases/silva-euk-28s-database-id98.fasta

Time to load the Burst trie: 2.0302s  
Begin parallel traversal ...  
Time of parallel traversal of automata: 19.0678s

Total number of reads found in current section: 95206

Time to output reads to file: 1.7558s  
Total number of reads found: 95206

The option '--log bilan' will create an overall statistics file called bilan.log,

> cat bilan.log

Time and Date

Settings:

r = 0.15

L = 18

reads file: ~/SRR106861-filtered.fasta

Results:

total reads: 105873

non-rRNA: 10667

rRNA: 95206

```
% rRNA: 89.92%
```

```
silva-bac-23s-database-id98.fasta 64.39%
silva-bac-16s-database-id85.fasta 25.53%
silva-euk-28s-database-id98.fasta 0.009445%
```

Now, we can manually check the number of reads matched per database:

```
> grep -c '>' rrna.*
```

```
rrna.silva-bac-16s-database-id85.fasta:27026
rrna.silva-bac-23s-database-id98.fasta:68170
rrna.silva-euk-28s-database-id98.fasta:10
```

#### 4.2.2 Example 3: sortmerna on multiple databases, with output of accepted reads directed to one single file

```
> sortmerna -n 3
 --db ~/sortmerna/rRNA_databases/silva-bac-23s-database-id98.fasta
 ~/sortmerna/rRNA_databases/silva-bac-16s-database-id85.fasta
 ~/sortmerna/rRNA_databases/silva-euk-18s-database-id95.fasta
 --accept rrna
 --454 SRR106861-filtered.fasta
 --log bilan
 -a 3
 -v
```

WARNING: option '--other' has been left blank, no output file for rejected reads ..

```

Welcome to SortMeRNA!
Copyright (C) 2012 Bonsai Bioinformatics Research Group
LIFL, Université Lille 1, CNRS UMR 8022, INRIA 2012

```

```
The size of the reads file <33862846> bytes
will be executed in 1 partial section(s) of size
<33862846> bytes
```

```
[Partial section # 1]

```

```
Time to mmap reads and set up pointers: 0.2184s
```

```
Begin analysis of: ./rRNA_databases/silva-bac-23s-database-id98.fasta
```

```
Time to load the Burst trie: 1.4594s
```

```
Begin parallel traversal ...
```

```
Time of parallel traversal of automata: 14.5262s
```

```
Begin analysis of: ./rRNA_databases/silva-bac-16s-database-id85.fasta
```

```
Time to load the Burst trie: 2.2632s
Begin parallel traversal ...
Time of parallel traversal of automata: 17.3602s
```

```
Begin analysis of: ./rRNA_databases/silva-euk-28s-database-id98.fasta
```

```
Time to load the Burst trie: 1.9980s
Begin parallel traversal ...
Time of parallel traversal of automata: 19.4148s
```

```
Total number of reads found in current section: 95206
```

```
Time to output reads to file: 0.9978s
Total number of reads found: 95206
```

Now, we can manually check the number of reads matched for all databases:

```
> grep -c '>' rrna.*
```

```
rrna.fasta:95206
```

#### 4.2.3 Example 4: sortmerna on paired-end reads (1 input file - paired-end reads are interleaved)

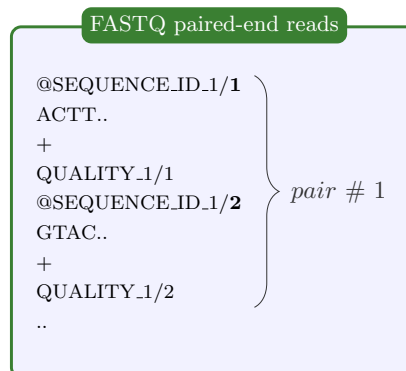


Figure 2: Paired-end read format accepted by SortMeRNA

This example illustrates three cases of output for paired-end reads: default, `--paired-in` and `--paired-out`.

We use the `set6-database.fasta` found under Section 3.2 of the Supplementary file (also available online at [bioinfo.lifl.fr/RNA/sortmerna/material.php](http://bioinfo.lifl.fr/RNA/sortmerna/material.php)). As for the reads, 5000 Illumina

paired-end reads were simulated using MetaSim on `set6-database.fasta`. The reads were then arranged to have 2475 pairs of rRNA, and the other 25 pairs where exactly one read is rRNA and the other is not (this is possible if one of the reads covers a low complexity region on the rRNA sequence).

**Remark:** The statistics in the `--log` file will always give the true number of reads classified as rRNA.

**Case 1: We don't care to keep the paired-end order in the output (default).**

```
> sortmerna -n 1
 --db set6-database.fasta
 --I paired-end-5000-reads.fasta
 --accept rrna
 --other nonrrna
 --log bilan
 -a 1
```

```
> cat bilan.log
```

```
Results:
total reads: 5000
non-rRNA: 50
rRNA: 4950
% rRNA: 99%
```

```
set6-database.fasta 99%
```

```
> grep -c '>' rrna.fasta nonrrna.fasta
rrna.fasta: 4950
nonrrna.fasta: 50
```

**Case 2: We want to accept all pairs with at least one read that hits (`--paired-in`).**

```
> sortmerna -n 1
 --db set6-database.fasta
 --I paired-end-5000-reads.fasta
 --accept rrna
 --other nonrrna
 --log bilan
 --paired-in
 -a 1
```

```
> cat bilan.log
```

```
Results:
total reads: 5000
non-rRNA: 50
rRNA: 4950
```

```

% rRNA: 99%

set6-database.fasta 99%

> grep -c '>' rrna.fasta nonrrna.fasta
rrna.fasta: 5000
nonrrna.fasta: 0

```

**Case 3:** We want to reject all pairs with at least one read that hits (`--paired-out`).

```

> sortmerna -n 1
 --db set6-database.fasta
 --I paired-end-5000-reads.fasta
 --accept rrna
 --other nonrrna
 --log bilan
 --paired-out
 -a 1

> cat bilan.log

Results:
total reads: 5000
non-rRNA: 50
rRNA: 4950
% rRNA: 99%

set6-database.fasta 99%

> grep -c '>' rrna.fasta nonrrna.fasta
rrna.fasta: 4900
nonrrna.fasta: 100

```

#### 4.2.4 Example 5: sortmerna on forward-reverse paired-end reads (2 input files)

SortMeRNA accepts only 1 file as input for the reads. If a user has two input files, in the case for the forward and reverse paired-end reads (see Figure 3), they may use the `merge-paired-reads.sh` script found in '`sortmerna/scripts`' folder to interleave the paired reads into the format of Figure 2.

The command for `merge-paired-reads.sh` is the following,

```
> bash ./merge-paired-reads.sh forward-reads.fastq reverse-reads.fastq outfile.fastq
```

Now, the user may input `outfile.fastq` to SortMeRNA for analysis.

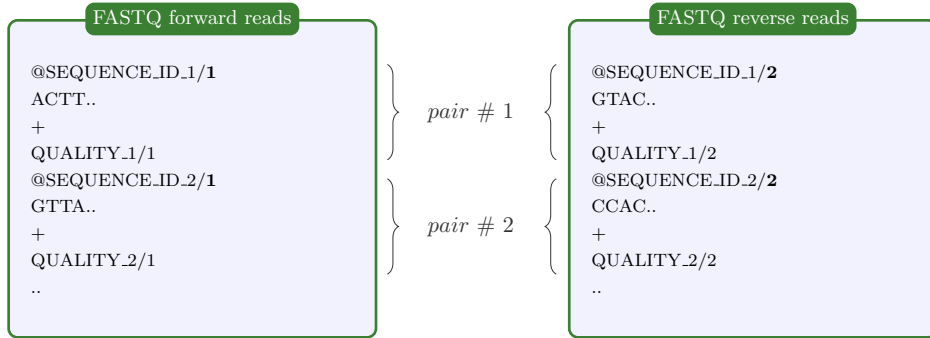


Figure 3: Forward and reverse reads in paired-end sequencing format

## 5 SortMeRNA parameters

There are two parameters in SortMeRNA which the user can moderate for performance: The length of the sliding window  $s$  (the seed), and the threshold ratio  $r$  of matching windows to the rRNA database for a read to be accepted. Both of these parameters and their default values are discussed in detail in *Section 2: Parameter Setting* of the supplementary file. The user may adjust the threshold parameter  $r$  with the sortmerna command-line option `-r [new ratio]`. To adjust the length of the sliding window  $s$ , the user must provide the option `-L [new length (even integer)]` when indexing an rRNA database using the `buildtrie` executable.